View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Queensland University of Technology ePrints Archive

Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

McCool, Christopher, Sa, Inkyu, Dayoub, Feras, Lehnert, Christopher, Perez, Tristan, & Upcroft, Ben
(2016)
Visual detection of occluded crop: For automated harvesting. In
IEEE International Conference on Robotics and Automation (ICRA 2016),
16-21 May 2016, Stockholm, Sweden.

This file was downloaded from: https://eprints.qut.edu.au/94274/

© Copyright 2016 [Please consult the author]

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

Visual Detection of Occluded Crop: for automated harvesting

Christopher McCool[†*], Inkyu Sa[†*], Feras Dayoub[*], Christopher Lehnert[*], Tristan Perez[*], and Ben Upcroft[*]

***Abstract*— This paper presents a novel crop detection system applied to the challenging task of field sweet pepper (capsicum) detection. The field-grown sweet pepper crop presents several challenges for robotic systems such as the high degree of occlusion and the fact that the crop can have a similar colour to the background (green on green). To overcome these issues, we propose a two-stage system that performs per-pixel segmentation followed by region detection. The output of the segmentation is used to search for highly probable regions and declares these to be sweet pepper. We propose the novel use of the local binary pattern (LBP) to perform crop segmentation. This feature improves the accuracy of crop segmentation from an AUC of 0.10, for previously proposed features, to 0.56. Using the LBP feature as the basis for our two-stage algorithm, we are able to detect 69.2% of field grown sweet peppers in three sites. This is an impressive result given that the average detection accuracy of people viewing the same colour imagery is 66.8%.**

I. Introduction

Current crop harvesting in horticulture crops is labor intensive, time-consuming, and not scalable under the increasing demands of food productivity. One way to tackle this problem is the deployment of farm robotics. To make robotic crop harvesting a reality, three key challenges have to be solved: automatic crop detection, automatic crop localisation, and automatic crop manipulation.

1) Accurate crop detection is essential to be able to detect the presence of a crop. This problem remains unsolved due to challenging conditions such as illumination variation and high levels of occlusion.
2) Crop localisation provides the exact location and orientation of a crop so that the best position for manipulation can be determined.
3) Crop manipulation and picking consists of being able to detach the crop without harming either the crop or plant.

In this paper, we address the first challenge of crop detection in a typical sweet pepper (capsicum) farm environment. Sweet pepper is chosen as it presents a range of challenges including varying crop colour (red and green for our experiments) as well as high levels of occlusion, as can be seen in Figure 1. Furthermore, as the crop is picked even when green (the same colour as the background) differentiating the crop and background (leaves) is extremely challenging.

We propose a crop detection system that relies on accurate pixel-level crop segmentation; the segmentation approach was initially presented in [1]. Inspired by Hung et al. [2], crop segmentation is performed using a conditional random field (CRF) and visual texture features. We propose the novel use of local binary pattern (LBP) features for crop segmentation and empirical evaluations show that this feature leads to impressive performance improvements, outperforming previous state-of-the-art features. Finally, our detection system uses the probability map from crop segmentation to search for highly probable regions and declares these as a detected crop. This provides robustness to occlusion by not assuming a particular shape on the visible crop.

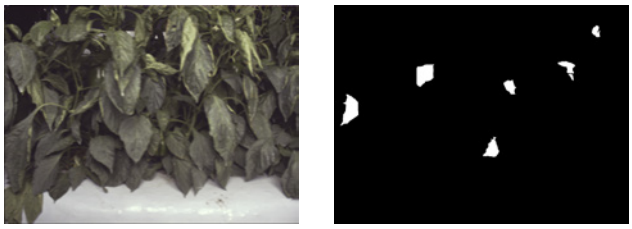


Fig. 1. An image of field grown sweet peppers highlighting the high level of occlusion and its similarity to the background. Left is a colour image. Right is the ground truth where sweet pepper are highlighted in white.

Developing this novel system leads to two major contributions:

- Develop a crop detection system which finds 69.2% of a highly occluded crop (sweet pepper). We show empirically that this is similar to the performance of humans who have an average accuracy of 66.8%.
- Propose the novel use of local binary patterns (LBPs) for crop segmentation which considerably outperforms previously proposed features.

As the crop is picked even when green (the same colour as the background) differentiating the crop and background (leaves) is extremely challenging (see the demonstration video[1]). In addition, we propose a novel evaluation metric based on a Bayesian framework that allows us to describe the uncertainty. We use this, and other standard metrics, to perform extensive evaluations using data collected from three commercial sites acquired during day and night. This dataset, along with the protocols and annotated ground truth imagery, is distributed to encourage further research in the area[1].

The remainder of the paper is structured as follows: Section II reviews prior work in horticultural crop segmentation and detection. Section III describes pixel- and region-level segmentation methods. Dataset acquisition and evaluation metrics are presented in the section IV. Section V describes experimental results and analysis for the proposed methods. Conclusions are drawn in section VI.

∗This research was supported by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016), Queensland University of Technology (QUT), Brisbane, Australia. c.mccool@qut.edu.au, i.sa@qut.edu.au

† Indicates equal contribution.

[1]The dataset and demonstration video is available at: goo.gl/T6djo0

II. Related Work/Background

Robotic crop harvesting and the methods for segmenting and detecting crops have been explored by several researchers [3], [4], [5], [6], [7], [2] and an overview of the field was provided in [8]. The grape detector of Nuske et al. [3], [4], initially presented in 2011, was one of the earliest crop detection systems. They detected grapes in an image based on a radial symmetry transform and then used this information to perform accurate yield estimation. A limitation of their system was that they could not detect partially occluded grapes (crops). However, as they were performing yield estimation and not accurate crop detection they were able to cope with this limitation.

Wang et al. [6] examined the problem of apple detection so that they could perform yield prediction. They developed a system that detected apples based on their colour and distinctive specaular reflection pattern. Further information, such as the the average size of apples, was used to either remove erroneous detections or to split regions that could contain multiple apples. Another heuristic employed was to accept as detections only those regions which were mostly round.

In 2013, Bac et al. [7] proposed a segmentation approach for sweet peppers and Hung et al. [2] proposed the use of conditional random fields for almond segmentation. Bac et al. aimed to develop a robotic harvesting system and so proposed a $C = 5$ class segmentation approach in order to build an accurate obstacle map. They used a six band multi-spectral camera (with bandwidths of between 40-60 nm^2) and used a range of features including the raw multi-spectral data, normalized difference indeces, as well as entropy based texture features. Experiments in a highly controlled glasshouse environment showed that this approach produced reasonably accurate segmentation results, however, the authors noted that it was not accurate enought to build a reliable obstacle map.

Hung et al. developed an almond segmentation approach in order to perform yield estimation. They proposed a $C = 5$ class segmentation approach which learnt features using a sparse auto-encoder (SAE). These features were then used within a CRF framework and was shown to outperform previous work. They achieved impressive segmentation performance, but did not perform object detection. Furthermore, they noted that occlusion presented a major challenge.

More recently, Yamamoto et al. [5] performed tomato detection by first performing $C = 4$ class segmentation. Colour and shape features were used to train a classifier and regression trees (CART) classifier. This produced a segmentation map and grouped connected pixels into regions. Each region was declared to be a detection and to reduce the number of false alarms they trained a non-fruit classifier using a random forest.

An issue with all of the prior work is the inability to perform accurate crop detection in challenging conditions.

[2]The longpass filter, $>900\,\mathrm{nm}$, has a bandwidth of $100\,\mathrm{nm}$.

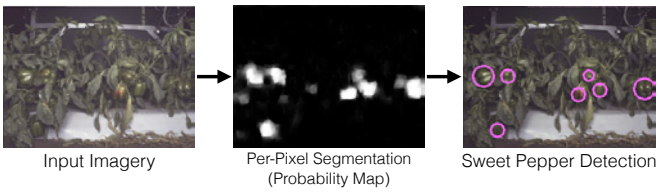


Fig. 2. An overview of the two stages for the proposed system using multi-spectral images as input. First, pixel-wise segmentation estimates the probability that a pixel belongs to the crop to produce a probability map $M(x, y)$. Second, we search the probability map to find highly probable regions and declare these as detected sweet pepper, represented by a magenta circle.

Several approaches have addressed only the crop segmentation task [7], [2] and not detection. Most of the work that has examined crop detection has predominantly been developed for yield estimation [3], [6] and so accurate detection was not necessary. The limited work that has examined accurate crop detection has done so for crops in controlled glasshouse environments [5]. As such the issue of crop detection in highly challenging conditions remains unsolved.

III. Proposed Approach

We propose a two-stage crop detection system that provides robustness to occlusion. First, pixel-level segmentation produces a probability map that a pixel belongs to the crop (sweet pepper). This per-pixel segmentation is robust to occlusion as it makes decisions for each pixel based only on a small region of the image; also unlike prior work [6] it does not include an explicit shape feature. Second, highly probable regions in the probability map are declared as being a detected crop. An example of these two stages is given in Figure 2.

A. Pixel-Level Segmentation

Inspired by Hung et al. [2], we perform crop segmentation using a CRF and visual texture features. We cast the problem as a $C = 2$ class segmentation of crop and not crop. This allows us to produce a probability map $M(x, y)$, similar to the one in Figure 2, that the crop occupies a particular pixel. This approach has the advantage of providing robustness against occlusion (since features are only taken from a small region) as well as minimising the amount of laborious annotation (as only the crop class needs to be annotated).

In contrast to Hung et al., we not only learn visual texture features but also explore the use of features which are robust to illumination. The three features considered are the SAE feature, similar to the one used by Hung et al., local binary patterns (LBPs) [9] and a histogram of gradients (HoG) [10]. We consider HoG and LBP features as they have been successfully used for other computer vision tasks such as object detection [10] and texture recognition [9]. While SAE features have previously been used for crop segmentation [2].

1) Sparse Auto-Encoder (SAE) Feature: An auto-encoder [11] is an unsupervised feature learning approach based on neural networks. The objective is learn a D-dimensional representation which can well represent the

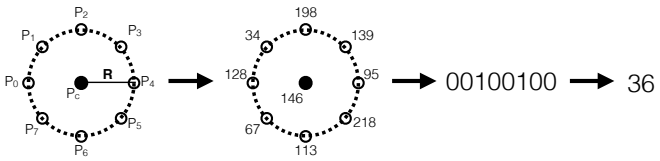


Fig. 3. An overview of local binary patterns are encoded. Using $P = 8$ sampling points on the circle of radius R the local binary pattern values are evaluated and represented as a binary string. This binary string can then be converted to an integer value.

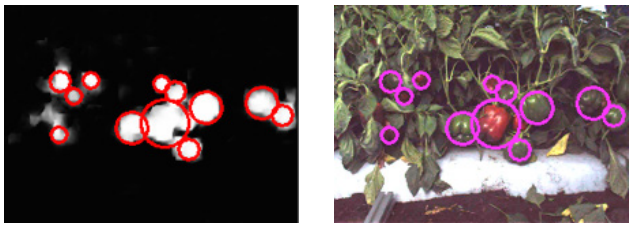


Fig. 4. Left is the probability map $M(x, y, \sigma)$ indicating which pixels, for the image on the right, is most likely to be a sweet pepper. Circles in both images are the detected sweet peppers using the LoG region detector.

input data $\mathbf{x}$ of dimension $D_{\mathbf{x}}$, where $D \ll D_{\mathbf{x}}$. This is achieved by learning a symmetric neural network which, in the first instance, progressively reduces the number of neurons until the middle layer conists of D neurons; this projects the input into a non-linear low-dimensional space. Layers after this middle layer then increase in size until it is the same size of the input data; this backprojects the low-dimensional representation. Thus, the objective is that the output of the network $\mathbf{y}$ matches the input of the network $\mathbf{x}$. This allows it to be trained in an unsupervised manner.

2) Histogram of Oriented Gradients (HoG): The HoG descriptor has been widely used for object detection [10]. The fundamental idea of the HoG feature is that an object can be described by the distribution of local gradients. These features are obtained by first, dividing an input image into patches (cells) of size $B_c \times B_c$ pixels. For each cell, a histogram of edge orientations are calculated and accumulated. Second, contrast-normalisation is performed in order to cope with illumination changes. A sampling `block` is then applied over the image region, in an overlapping manner, to accumulate the histograms from $B_b \times B_b$ cells[3].

3) Local Binary Pattern (LBP) feature: The local binary pattern is a simple and powerful feature that has shown impressive performance for several computer vision tasks including image, video, face and texture recognition [12], [9], [13]. It is both computationally efficient and robust to illumination variations as it is computed by performing a set of pixel comparisons, illustrated in Figure 3. The pixel of interest (central pixel) P_c is compared to the P surrounding pixels of radius R resulting in a code given by,

$$LBP_{\text{P,R}} = \sum_{p=0}^{P-1} h\left(P_c, P_p\right) 2^p, \tag{1}$$

where P_p is the value of the p-th pixel and

$$h\left(x, y\right) = \begin{cases} 1 & \text{if } y \geq x \\ 0 & \text{if } y < x \end{cases} \tag{2}$$

is a thresholding function.

B. Crop detection

To detect crop regions we use a Laplacian of Gaussian (LoG) multi-scale blob detector [14] on the probability map $M(x, y)$ generated from the previous step. The advantage of this approach is that groups surrounding pixels together provided they are of sufficiently high probability they will be grouped together. This means that it retains some of the robustness to occlusion from the previous step, which is one of our aims.

The multi-scale LoG blob detector searches the probability map $M(x, y)$ by applying a Gaussian kernel $\mathcal{G}(x, y, \sigma_n)$ over the image. The size of the Gaussian kernel is dictated by σ_n and by using N values it becomes a multi-scale approach. When the n-th Gaussian kernel is applied to the image a particular response $\mathcal{L}(x, y, \sigma_n)$ is produced. These responses are collated to form $\mathcal{L}(x, y, \sigma_{1,\ldots,N})$ and we can compute the Laplacian of this $\nabla^2\mathcal{L}(x, y, \sigma_{1,\ldots,N})$. Regions are then found by finding the local maxima of $\nabla^2\mathcal{L}(x, y, \sigma_{1,\ldots,N})$.

There are four key parameters; σ_{min} and σ_{max} are the minimum and the maximum scale value for the Gaussian kernel, N is the total number of scales, and τ_r is the threshold for region detection. Local maxima smaller than τ_r are discarded. Given ground truth images of sweet pepper, we empirically choose these values.

Using this approach we produce sweet pepper detection results such as those presented in Figure 4.

IV. Experimental Data and Protocols

We evaluate our crop detection system on sweet pepper grown on two farms. Sweet pepper presents several challenges including varying crop colour (red and green for the farms chosen) in addition to high levels of occlusion. An example of this is provided in Figure 5.

A. Data Acquisition

To acquire the data we used an acquisition system that was able to mount a single row of sweet pepper plants, shown in Figure 6. A multi-spectral camera, the JAI AD-130GE, was used to acquire imagery and a Novatel GPS was used to record accurate GPS data; the GPS data was not used in this work. To minimise the effect of illumination variation, a canopy was placed around the cart and a set of LED (visible and NIR) lights was mounted behind the camera. An example of the multi-spectral imagery acquired is given in Figure 5.

B. Datasets and Evaluation Protocols

In total, we acquired three datasets of field grown sweet peppers. The datasets are referred to as $G1$, $G2$, and $S1$. The prefix, G or S, refers to a particular farm and the number refers to the acquisition number. Two sets of data were acquired from one farm (G) with a delay of 4 months between acquisition, also, the data was acquired from a new

[3]The common values are $B_c = 8$, $B_b = 2$ with an overlap of 50%.

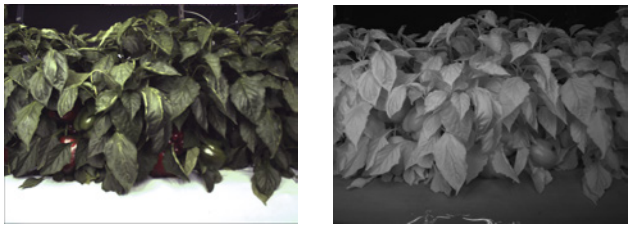


Fig. 5. Example of NIR (left) and visible (right) spectrum image. Sweet pepper provide a clear response in the NIR spectral range while the visible spectral range provides useful information such as colour.

set of sweet peppers which were grown in a new area of the farm.

Each dataset consists of acquisition during the afternoon (day) and the evening (night). This was done so that we could investigate the impact of illumination on system performance. Even though a canopy was used to reduce the impact of daylight illumination, there was still a considerable amount of light that filtered through. For this reason we refer to the day data as semi-controlled illumination and the night data as controlled illumination.

A subset of 103 images was manually annotated to highlight where sweet pepper are visible and used to evaluate our systems. A summary of the number of images is given in Table I.

TABLE I

NUMBER OF ANNOTATED IMAGES FOR EACH DATASET.

	Day/Night	Train	Valid.	Test	Num. Images
G1	**Day**	4	4	3	11
	Night	7	7	6	20
S1	**Day**	3	3	3	9
	Night	6	6	6	18
G2	**Day**	9	6	9	24
	Night	9	6	6	21

To evaluate our systems we divided the annotated data into training, validation and test sets. The training set was used to train the models such as the CRF and auto-encoder. The validation set was used to choose the optimal system as well its parameters and a threshold τ which was then applied to the test set. The test set was then used to present the final performance of the system, neither the parameters nor the threshold τ were tuned on this set. Using these three splits allows us to separate the optimisation of system parameters from the final evaluation of system performance. This differs to much of the previous work which optimised their system, parameters and thresholds on their test set and not on a separate validation set.

C. Performance Measures

The performance of our systems are presented using standard measures as well as a novel probabilistic approach. We provide a probabilistic approach to interpret the results as this allows us to describe our uncertainty about the performance of the system. Below, we first describe the standard measure that we use to evaluate the performance of our system and then present our novel probabilistic approach.

1) Evaluation of Segmentation - per pixel: To evaluate the performance of segmentation we use the area under the

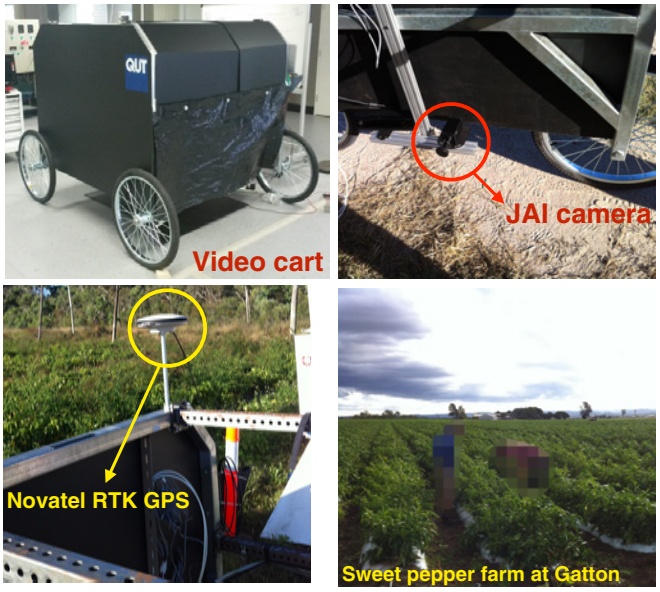


Fig. 6. QUT video cart developed for sweet pepper data collection [15].

curve (AUC) of the precision-recall curve. Precision (P) and recall (R) are given by,

$$P = \frac{T_p}{T_p + F_p}, \qquad R = \frac{T_p}{T_p + F_n}, \tag{3}$$

where T_p is the number of true positives (*correct detections*), F_p is the number of false positives (*false alarms*), and F_n is the number of false negatives (*mis-detections*).

2) Evaluation of Crop Detection: We evaluate the performance of the detection system using the detection rate (DR) and the average false positive rate per image (aFPI). These are then plotted for varying thresholds as an ROC curve. The DR and aFPI are given by

$$DR = \frac{T_p}{N_T}, \qquad aFPI = \frac{F_p}{N_I}, \tag{4}$$

where T_p is the number of true positives, N_T is the total number of crop (sweet pepper) regions, F_p is the number of false positives, and N_I is the number of images in the set. These measures are used for detection, rather than precision recall, as they can be directly related to a physical quantity (the number of crop detected).

D. Predictive Probability of Successful Detection

Although measures such as the the area under the precision-recall curve and detection rate provide a figure of merit, the difference between scores is hard to interpret if we would like to compare algorithms. In addition, there is no characterisation of uncertainty of the results. Hence, we propose a characterisation of performance based on probability theory, which we take as a description of our uncertainty about a hypothesis related to the performance of the algorithms. Understanding this uncertainty is a key aspect for the process of choosing a particular algorithm and also for making decisions given the uncertainty about its performance or reliability. We summarise our approach for the problem

of detection, but this is equally applicable to the $C = 2$ class segmentation problem.

The detection problem considers the presence, or lack thereof, of a target object (sweet pepper) in an image. We can consider two propositions (which can either be true or false):

$O = \{\text{Target capscium is present in the image}\}$;

$A = \{\text{Algorithm accuses the presence of the sweet pepper}\}$.

We then define the hypothesis or proposition of interest:

$$H = (A|O), \tag{5}$$

which states that the detection and classification algorithm provides the correct detection given that the sweet pepper is in the image. In relation to standard terminology, the proposition H is associated with a true positive (correct detection).

We propose as a metric of performance the predictive probability that H is true conditional on all evidence at hand; namely $P(H|D, B)$, where D is a proposition related to the data we use to test the algorithm and B is a proposition that summarises the background information.

In [16], we provide details on how $P(H|D, B)$ can be computed as a predicted probability using a Bayesian framework. We consider the data D as a sequence of Bernoulli trials, in which the algorithm either succeeds of fails with R number of successes over N trials. This gives the likelihood model $p(D|\theta, B) = \theta^R(1-\theta)^{N-R}$, where $0 \leq \theta \leq 1$ is a parameter. From Bayes Theorem we can compute the posterior distribution $p(\theta|D, B)$.

If we adopt as a measure of performance, the predictive probability of success in the next trial, then the sought predicted probability–obtained by marginalisation–is the posterior mean [16]:

$$P\left(H|D, B\right) = \int_0^1 \theta\, p(\theta|D, B)\, d\theta. \tag{6}$$

If these calculations are made based on a uniform prior $p(\theta|B)$ (motivated by maximum entropy), then $P\left(H|D, B\right) = (R+1)/(N+2)$, where for the detection example $N = N_T$ and $R = T_p$. Hence, we call $P(H|D, B)$ the *Predictive Probability of Successful Detection*. The proposed metric provides a novel way of assessing performance whilst capturing uncertainty.

We note the other priors can be used if we have access of further information about the algorithms being tested. We also note that the similar analysis can be conducted for misdetection by a re-definition of the propositions A and O above. See [16] for details.

V. Experimental Results

Three sets of experiments are performed to evaluate the effectiveness of our proposed detection system. The first set of experiments analyses the segmentation performance along with the impact of capturing data in controlled and semi-controlled illumination conditions; as mentioned in Section IV-B we refer to the day data as the semi-controlled illumination and the night data as controlled illumination. The second set of experiments analyses the detection performance and an example video of our proposed system detecting capsicum along 290 m of a commercial sweet pepper farm is available[1]. The third set of experiments examines the issue of occlusion. For all of our experiments we use an open source CRF implementation [17].

TABLE II

Segmentation performance in terms of AUC.Highlighted in bold is the best performing system.

	Day (Semi-Controlled)		Night (Controlled)	
	Valid.	**Test**	**Valid.**	**Test**
Combined+Colour	**0.62**	**0.58**	**0.70**	**0.68**
Combined	0.57	0.53	0.64	0.60
LBP	0.57	0.51	0.61	0.56
Baseline (SAE)	0.08	0.11	0.21	0.10
HoG	0.03	0.05	0.16	0.15

For comparison, we use auto-encoder features similar to those used by Hung et al. [2], however, when using these features we use just the NIR imagery[4]. This is because we incorporated colour as a separate feature using the HSV colour space for each pixel. It was not possible to compare to the prior work on sweet pepper segmentation conducted by Bac et al. [7] due to their use of a specialised six band multi-spectral camera.

A. Experiment I: sweet pepper segmentation (day and night)

In this experiment we analyse the effectiveness of three visual texture features along with the impact of using data captured during either the day or night. The visual features are extracted from NIR imagery, which is more consistent than the colour imagery, and we present the results in Table II.

Results in Table II show that the LBP is a robust and effective feature to use for crop segmentation. It provides considerably higher accuracy and precision than the Baseline (SAE) and HoG features with an AUC of 0.56 compared to 0.10 and 0.15 for the Baseline and HoG features respectively. However, despite the performance difference between these features combining them leads to further improvements.

The combination of all of the visual features (Combine) provides a relative performance improvement, for the AUC, of 7%. Incorporating colour as a feature, using the HSV colour space, provides further improvements with an AUC of 0.68. This is a relative improvement of 21% compared to using just the LBP feature and highlights the importance of using a variety of features to cope with this challenging problem.

Finally, it can be seen that using the night data (controlled illumination) always leads to improved performance. For example, using night data provides a relative improvement

[4]It was not possible to compare directly to the system of Hung et al. [2] as they did not provide their parameters.

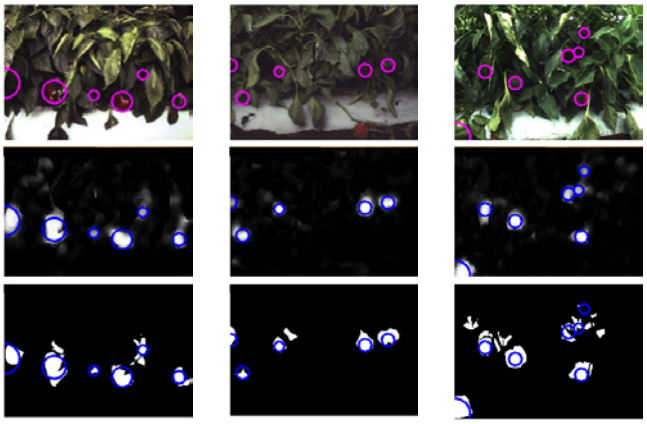


Fig. 7. Three example images from the test set. The top row shows colour image with identified regions superimposed in purple. The middle row shows segmentation results with identified regions superimposed in blue and the bottom row shows ground truth with identified regions superimposed in blue.

TABLE III

TABLE OF DETECTION PERFORMANCE IN TERMS OF DETECTION RATE (DR), PREDICTIVE PROBABILITY OF SUCCESSFUL DETECTION (PPSD), AND AVERAGE FALSE ALARMS PER IMAGE (AFAI).

	DR	**PPSD**	**aFAI**
Validation set	75.8%	0.75	1.2
Test set	69.2%	0.69	2.1

of 10% when using the LBP and 13% for the Combined system. An even larger relative improvement is obtained for the system that uses colour information with a relative improvement in AUC of 17%. We attribute this improvement in performance to being able to better control the illumination. As such, we recommend where possible to acquire imagery during night as this has a considerable impact upon performance.

B. Experiment II: crop (sweet pepper) detection

In this experiment we analyse the performance of our proposed crop (sweet pepper) detection system. Based on experiments on the validation set we found the optimal parameters for detection. These parameters include the threshold τ for probable regions, size of the multi-scale Laplacian of Gaussians and also an appropriate minimum size of the crop to detect[5].

The results in Table III show that our system has a DR of 69.2% and PPSD of 0.69 for field grown sweet peppers. Examples of the detection results, on the test set, which highlight the difficulty of the problem are given in Figure 7 and a video of results over more of the test data is also available[1]

To put the performance of our system in perspective, we compare its performance against that of an untrained person who viewed the colour imagery. The participants are referred to as untrained as they are not trained sweet pepper pickers, they were asked to click on any area in the image which corresponded to a sweet pepper and were given one minute

[5]On the validation set we found that an area of 100 pixels was appropriate and this corresponds to just 0.2% of the entire image area.

Fig. 8. An overview of the testing environment. The camera on the arm is moved across a set trajectory to capture images of the sweet pepper with varying levels of occlusion.

to detect all of the sweet pepper in a single image. In total there were seven participants.

The results in Table IV show the average, best, and worst performance of the participants. In terms of the DR and PPSD, the performance of our algorithm is comparable to that of the participants—better than the worst participant but below the best participant. This demonstrates that our proposed system provides a highly competitive, and automatic, method to detect field grown sweet pepper.

An issue with our system is that it does result in a number of false alarms. Our proposed system has an aFAI of 2.1, which is higher than for the participants who had an average aFAI of 1.2. We believe that the number of false alarms produced by our system could be reduced if we exploit temporal information or use multiple images of the same area before declaring a region as belonging to the crop (sweet pepper).

TABLE IV

PERFORMANCE OF HUMAN PARTICIPANTS ON THE TEST SET.

	DR	**PPSD**	**aFAI**
Average participant performance	66.8%	-	1.2
Most accurate participant	76.5%	0.76	1.3
Least accurate participant	60.7%	0.61	1.2

C. Experiment III: A study of correlation between occlusions and detection performance

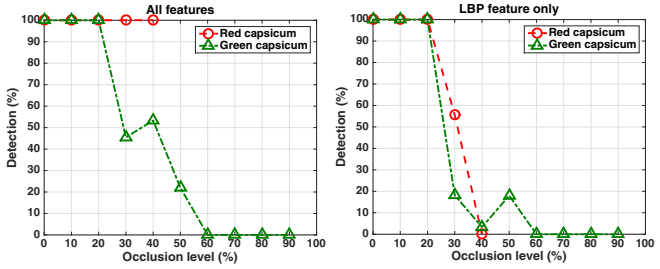
To examine the robustness of our system to varying levels of occlusion we performed a set of controlled experiments where the camera was mounted on a robotic arm (UR5); the same camera and illumination setup was used.The arm moved along a preset trajectory so that different views of the sweet pepper (attached to the plant) were obtained, this provided us with varying levels of occlusion due to changes in view point. An overview of the experimental setup is given in Figure 8.

Using the above experimental setup, we acquired images for a red and green sweet pepper with varying levels of occlusion. In total 197 images were acquired for each sweet pepper. This data was then manually annotated so that we could calculate the percentage of occlusion

$$OC_{k,i} = (1 - O_{k,i}/O_k) \times 100\%, \quad (7)$$

where $O_{k,i}$ is the number of pixels visible for the i-th image of the k-th crop and O_k is the total number of pixels that

Fig. 9. Example imagery captured for the red (top row) and green (bottom row) sweet peppers. The red sweet pepper has 36% of its visible area occluded while the green sweet pepper has 83% occluded.



(a) (b)

Fig. 10. Occlusion level versus detection rate for two feature combinations. The performance of our system that uses All features (Combined+Colour) is given in (a) and a system using only the LBP features can be seen in (b).

could be visible for the k-th crop. The reference value O_k was estimated manually based on viewing the entire footage.

Using our pre-trained model, from the previous sections, we evaluated the detection rate with varying levels of occlusion. The results in Figure 10(a) show that, irrespective of colour, our system can detect all instances where the crop is occluded by 20% or less. For greater levels of occlusion detection performance for green sweet pepper drops considerably, with the system being unable to find sweet pepper with 60% or more occlusion. However, our system is able to detect all instances of red sweet peppers[6].

We perform the same analysis with a system using only the LBP features. It can be seen in Figure 10(b) that when there is 20% or less occlusion a system using just LBP features can detect all instances of sweet pepper. However, the detection rate degrades considerably once higher levels of occlusion are encountered and is much lower than the combined system (Combined+Colour). This suggests that the proposed multi-feature approach provides a level of robustness to occlusion.

VI. Summary and Future Work

We present a novel sweet pepper (capsicum) vision-based detection system which can find 69.2% of sweet peppers in real-world farms. This is an impressive result given that humans have similar performance on the same colour imagery, on average they detect 66.8% of sweet peppers.

In developing this system, we propose a novel crop segmentation system that outperforms previously proposed features. The novel use of LBP features for crop segmentation provides considerable improvements over SAE features, similar to those proposed by Hung et al., and also HoG features. Furthermore, the combination of visual features leads to further improvements leading to a crop segmentation system with an AUC of 0.68 in challenging conditions.

Future work will consider ways to incorporate temporal information to improve the detection rate and reduce the number of false alarms. Finally, work should examine methods to learn appropriate features using deep learning techniques.

[6]The maximum occlusion for red sweet pepper was approximately 40%.

References

[1] I. Sa, C. McCool, C. Lehnert, and T. Perez, "On visual detection of highly-occluded objects for harvesting automation in horticulture," in *Workshop on Robotics in Agriculture at the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[2] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5314–5320, Nov 2013.

[3] S. T. Nuske, S. Achar, T. Bates, S. G. Narasimhan, and S. Singh, "Yield estimation in vineyards by visual grape detection," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '11)*, September 2011.

[4] S. Nuske, K. Wilshusen, S. Achar, L. Yoder, S. Narasimhan, and S. Singh, "Automated visual yield estimation in vineyards," *Journal of Field Robotics*, vol. 31, no. 5, pp. 837–860, 2014.

[5] K. Yamamoto, W. Guo, Y. Yoshioka, and S. Ninomiya, "On Plant Detection of Intact Tomato Fruits Using Image Analysis and Machine Learning Methods," *Sensors*, 2014.

[6] Q. Wang, S. T. Nuske, M. Bergerman, and S. Singh, "Automated crop yield estimation for apple orchards," in *13th Internation Symposium on Experimental Robotics (ISER 2012)*, no. CMU-RI-TR-, July 2012.

[7] C. W. Bac, J. Hemming, and E. J. Van Henten, "Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper," *Comput. Electron. Agric.*, vol. 96, pp. 148–162, Aug. 2013.

[8] K. Kapach, E. Barnea, R. Mairon, Y. Edan, and O. Shahar, "Computer vision for fruit harvesting robots: state of the art and challenges ahead," *International Journal of Computational Vision and Robotics*, vol. 3, pp. 4–34, 2012.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893 vol. 1, June 2005.

[11] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, 2011.

[12] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen, "Rotation-invariant image and video description with local binary pattern features," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1465–1477, 2012.

[13] T. Ahonen, A. Hadid, and T. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2037–2041, 2006.

[14] T. Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention," *International Journal of Computer Vision*, vol. 11, no. 3, pp. 283–318, 1993.

[15] C. McCool, C. Lehnert, D. Hall, B. Upcroft, and T. Perez, "Queensland DAFF Strategic Investment in Farm Robotics (SIFR) Milestone Report - Studies on Weed Identi cation and Weed Destruction," tech. rep., QUT, 2015.

[16] T. Perez, I. Sa, C. McCool, and C. Lehnert, "A bayesian framework for the assessment of vision-based weed and fruit detection and classification algorithms," in *Workshop on Robotics in Agriculture at the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[17] J. Domke, "Learning Graphical Model Parameters with Approximate Marginal Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2454–2467, 2013.